Executive Steering Committee
For A.C.E. Policy II
(ESCAP II)
Report 4
September 21, 2001

ESCAP II: A.C.E. Erroneous Enumerations Errors: Analysis of Census Discrepant Persons

Elizabeth A. Krejsa
Planning, Research, and
Evaluation Division

U S C E N S U S B U R E A U
Helping You Make Informed Decisions

EXECUTIVE SUMMARY

Is there an underestimate of census discrepant errors?

As shown by the discussion below, the net effect of erroneously identifying discrepant persons as correct enumerations in production and vice versa is a difference of 5,811 too many correct enumerations in production, with a standard error of 31,835. Therefore, this difference is insignificant. It is 0.002 percent of the 256,356,408 correct enumerations in the weighted production results.

Discrepant results are errors that could include falsification (the amount is uncertain), but do not include honest mistakes made by the interviewers or respondents. A person is classified as discrepant during the matching operation if three knowledgeable respondents indicate not knowing him or her in a followup interview (either the Evaluation Followup or Person Followup). The three knowledgeable respondents must have answered "No" to the questions,

- Do you know or have you heard of (the census person)?
- Do you know someone else who might know (the census person)?

The evaluation data used are a combination of results (referred to as the combined results) from the Matching Error Study, an independent rematch of the production data, and the Evaluation Followup Interview matching, an enumeration status coding operation based on the results of a personal visit reinterview. Both studies were conducted in 2,259 Accuracy and Coverage Evaluation clusters (an approximate 1-in-5 sample of the Accuracy and Coverage Evaluation clusters).

Discrepant Persons undetected in the production matching operation based on match code alone

While the production process identified 606,146 discrepant persons, 213,533 of those were erroneously identified as discrepant in the combined results. An additional 236,191 discrepant persons who were not identified in production as discrepant were identified in the combined results. Therefore, the net residual discrepant persons in the E-sample is 22,658 (0.008 percent of the E-sample people) with a standard error of 43,701.

Demographic Characteristics of the Combined Results Discrepant Persons

The largest discrepancy rate for tenure, age, type of enumeration area, race, and urbanicity was 0.11 percent. Thus, the error in identification of discrepant persons does not have much of an impact on any specific demographic group.

Effect on the estimate of correct enumerations

While the count of residual discrepant persons gives us insight into differences in matching and followup information obtained, this number does not adequately represent the impact on the dual system estimates (DSEs). The E-sample term in the DSE is the number of correct enumerations (CE) divided by the number of E-sample people (E). The misclassification of discrepant persons only affects the CE term, not the E term.

122,456 correct enumerations were identified in the combined results that had been reported as discrepant persons in production.

128,267 correct enumerations were identified in production that were reported as discrepant in the combined results.

The net effect of erroneously identifying discrepant persons as correct enumerations in production and vice versa is a difference of 5,811 too many CEs in production with a standard errof of 31,835. This difference is 0.002 percent of the 256,356,408 CEs in the weighted production results.

1. BACKGROUND

This report answers the question: How many discrepant people were undetected in the production matching operation?

Discrepant results are errors that do not include honest mistakes made by the interviewers or respondents and could be falsification but the amount is uncertain. A person is classified as discrepant during the production matching operation if three knowledgeable respondents indicate not knowing him or her in the Person Followup interview (PFU). The three knowledgeable respondents must have answered "No" to the questions,

- Do you know or have you heard of (the Census person)?
- Do you know someone else who might know (the Census person)?

Once a person is classified as discrepant he or she is flagged on the file. In missing data processing these people are assigned zero probability of correct enumeration (Childers, 2000).

2. METHODS

To determine the amount of discrepant persons that are not identified in production, we use the combined results of E-sample discrepant persons. The combined results are the best match codes resulting from two operations - the matching error study rematch and the Evaluation Followup interview (EFU).

2.1 The matching error study

The matching error study is a rematch of the production data in approximately 1-in-5 A.C.E. clusters. This sample, referred to as the evaluation clusters, consists of a total of 2,259 clusters. The same matching rules are followed in this study as in the production matching. Discrepancies between the production and rematch results are reviewed and reconciled by the expert clerical matching analysts (Bean, 2001).

2.2 Matching the EFU data

The EFU is a personal visit reinterview of people listed in either the census or the A.C.E. Person Interview. It was conducted in January and February, 2001 in the evaluation clusters. The purpose of the EFU is similar to the PFU in that it gathers information to resolve conflicts between A.C.E. and Census and to determine residence status. In addition, the EFU identifies reasons why a person may have been erroneously listed or not listed as a census day resident in the A.C.E. or Census.

After the EFU interview is complete, a clerical matching operation takes place. The EFU matching operation is similar to the production After Followup matching operation except the

matchers use the information from both the PFU interview and the EFU interview to determine true residence status on census day.

A person can be classified as discrepant in this operation if three knowledgeable respondents indicate not knowing him or her in the EFU interview, in the same way a person can be classified as discrepant in PFU.

3. LIMITS

Because the EFU interview takes place 9 months after Census Day there is concern that information reported in the EFU may not be as valid as the information received during the production process. For this reason, matchers are given the option to reject the information received in the EFU interview in favor of the production match code. Information obtained for 10.6 percent of people who were followed up in EFU was rejected. In general, if information is obtained about a person in the PFU interview but not in the EFU interview, the EFU interview for that person is rejected and the production match code is kept, as opposed to coding the person discrepant in the EFU (Green, 2001).

4. RESULTS

Table 1 below shows the count of matching outcomes between production and the combined results weighted to the E-sample. When three knowledgeable respondents indicate not knowing the followup person or if the name is found to be a pet, a matcher codes the E-sample person as discrepant in the block cluster. This means that the person may have existed, but should not have been enumerated in the census within this block cluster and thus was erroneously enumerated (Childers, 2000). The calculation of the production discrepant and non-discrepant persons is done only within the 2,259 evaluation clusters and are weighted to the national level. Standard errors are included in parenthesis.

Table 1. Comparison of Matching Outcome of Discrepant Persons from the Combined Results vs. Production

Combined Results	Production Results		
	Discrepant Person	**Non Discrepant Person**	**Total**
Discrepant Persons Weighted	392,612 (0.15% of weighted total) (56,739)	236,191 (0.09% of weighted total) (34,427)	**628,803 (0.23% of weighted total)** (69,597)
Non Discrepant Persons Weighted	213,533 (0.08% of weighted total) (29,314)	267,841,465 (99.68% of weighted total) (6,481,500)	**268,054,998** (6,483,067)
Total	**606,145 (0.23% of weighted total)** (68,509)	**268,077,656** (6,482,526)	**268,683,802** (6,486,463)

4.1 Discrepant Persons undetected in the production matching operation

While the production process identified 606,146 discrepant persons, 213,533 of those were erroneously identified as discrepant according to the combined results. An additional 236,191 discrepant persons who were not identified in production as discrepant were identified in the combined results. Therefore, the difference in identification of discrepant persons in the E-sample is 22,658 (0.008 percent of the E-sample people) with a standard error of 43,701.

4.2 Effect on the estimate of correct enumerations

While the count of residual discrepant persons gives us insight into differences in matching and followup information obtained, this number does not adequately represent the impact on the dual system estimates (DSEs). The E-sample term in the DSE is the number of correct enumerations (CE) divided by the number of E-sample people (E). The misclassification of discrepant persons only affects the CE term, not the E term.

4.2.1 How coding differences affects the estimate of correct enumerations

If we look more closely at the coding disagreements we can see that the effect of the disagreement on the CE count is differential by original enumeration status.

People erroneously enumerated in the census (for reasons such as geocoding errors, duplicates, and discrepant persons) are assigned a CE probability of zero. Therefore, people who were identified in production as erroneously enumerated for a reason other than being discrepant but were then identified as being discrepant in the combined results have no impact on the CE component of the DSE.

In addition, some people in production were coded as unresolved instead of discrepant only because three people were not contacted to confirm that the person was not known. When CE probabilities were imputed for the unresolved people, these possibly discrepant-unresolved people were given lower probabilities than other unresolved people. Thus, a change in match code to a discrepant person in these cases has a much lower impact on the CE component of the DSE compared to other unresolved cases.

4.2.2 Net effect of discrepant persons on the weighted correct enumeration count

Taking into account the CE probabilities results in a more accurate picture of the effect of the residual discrepant persons. Taking these into account:

- The combined results identified 122,456 CEs (with a standard error of 21,559) that had been reported as discrepant persons in production.

- Production reported 128,267 CEs (with a standard error of 24,057) that were reported as discrepant in the combined results.

- The net effect of erroneously identifying discrepant persons as CEs in production and vice versa is a difference of 5,811 CEs too many in production (with a standard error of 31,835). This difference is 0.002 percent of the 256,356,408 CEs in the weighted production results.

4.3 Demographic Characteristics of the Combined results Discrepant Persons

Because one common source of discrepant persons is an interviewer, discrepant persons may be clustered by geography as well as by stratification characteristics such as tenure, age, and sex. Such clustering could impact the DSEs in those subgroups.

4.3.1 Evidence of discrepancies by interviewer

Logic dictates that interviewers who enter discrepant results are most likely to do so for an entire household rather than for part of a household. Therefore, one indication that an interviewer entered discrepant results is a household in which every person is discrepant. Of the 377 (unweighted) households in the combined results that contained discrepant persons, 295 households (78 percent) were whole household discrepancies.

Another indication that an interviewer entered discrepant results is that the person was enumerated via an enumerator-filled questionnaire versus a self-reporting form. In the combined results, 663 of the 799 discrepant persons (83 percent) were from whole household discrepancies and enumerated via an enumerator-filled questionnaire. These 663 people are more likely the result of an interviewer entering discrepant information than are any of the remaining 136 people.

There were 186 clusters identified in the combined results which contained discrepant persons. Fourteen clusters contained more than five whole household discrepancies, and five of those contained more than ten whole household discrepancies. In the five clusters that contained more than ten whole household discrepancies, all of the discrepant people were reported on an enumerator-filled questionnaire. These households account for 18.3 percent of all discrepant households and 23.2 percent of all discrepant people in the combined results. This clustering affect, though small, may be significant if other types of errors are noted in the same clusters.

4.3.2 Poststratification characteristics

Discrepancies between production and the combined results do not appear to be clustered demographically. The largest discrepancy rate for tenure, age, type of enumeration area, race, and urbanicity was 0.11 percent. Thus, the error in identification of discrepant persons does not seem to have an impact on any specific demographic group. Discrepancies between production and the combined results are also not clustered geographically.

4.4 Quality of the Combined results Data

The differences between the combined results results and production are in the direction of what was expected. We expected that the EFU survey would produce additional information that would result in a change in match code. We also expected that a large number of unresolved cases would be coded as discrepant in the combined results. These cases were identified as possibly discrepant in production because three people were not contacted to confirm that the person was not known. In the EFU, however, three people were contacted and matchers were able to code these people discrepant.

Of the 213,533 erroneous discrepant persons, 23.7 percent are the result of matching error in production and 76.3 percent are the result of additional information obtained during the EFU interview that determined the person was not discrepant.

Of the 236,191 newly identified discrepant persons, 26.5 percent are the result of matching error in production and 57.6 percent are the result of conversion of unresolved match codes to discrepant matching codes based on the EFU. The remainder, 15.9 percent, are the result of conversion of other match codes to discrepant match codes based on the EFU.

5. CONCLUSIONS

The net effect of erroneously identifying discrepant persons as correct enumerations in production and vice versa is a difference of 5,811 too many correct enumerations in production, with a standard error of 31,835. Therefore, this difference is insignificant. It is 0.002 percent of the 256,356,408 correct enumerations in the weighted production results. The difference is not concentrated geographically or in any demographic group.

6. REFERENCES

Bean, Susanne. "Matching Error Study Clerical Matching Guide." PRED TXE/2010 Memorandum Series, Chapter CM-MESM-S-02-R01. February 6, 2001.

Coverage Measurement Staff. "Evaluations Operational Plan for the Census 2000 Accuracy and Coverage Evaluation." Draft Specification. July, 2000.

Childers, Danny R. "Accuracy and Coverage Evaluation: The Design Document." DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-01. October 11, 2000.

Green, Alicia. "Measurement Error Reinterview Matching Specification." PRED TXE/2010 Memorandum Series, Chapter CM-MER-S-03. February, 2001.